NREL 01-38

THERMAL TOLERANT EXOGLUCANASE FORM ACIDOTHERMUS CELLULOLYTICUS

By

William S. Adney
Shi-you Ding
Suzanne Lantz McCarter
Stephen R. Decker
Todd B. Vanzant
Michael E. Himmel

THERMAL TOLERANT EXOGLUCANASE FROM ACIDOTHERMUS CELLULOLYTICUS

Government Interests

The United States Government has rights in this invention under Contract No. DE-AC36-99G010337 between the United States Department of Energy and the National Renewable Energy Laboratory, a Division of the Midwest Research Institute.

Field of the Invention

The invention generally relates to a novel exoglucanase from Acidothermus cellulolyticus, Gux1. More specifically, the invention relates to purified and isolated Gux1 polypeptides, nucleic acid molecules encoding the polypeptides, and processes for production and use of Gux1, as well as variants and derivatives thereof.

Background of the Invention

Plant biomass as a source of energy production can include agricultural and forestry products, associated by-products and waste, municipal solid waste, and industrial waste. In addition, over 50 million acres in the United States are currently available for biomass production, and there are a number of terrestrial and aquatic crops grown solely as a source for biomass (A Wiselogel, et al. Biomass feedstocks resources and composition. In CE Wyman, ed. Handbook on Bioethanol: Production and Utilization. Washington, DC: Taylor & Francis, 1996, pp 105-118). Biofuels produced from biomass include ethanol, methanol, biodiesel, and additives for reformulated gasoline. Biofuels are desirable because they add little, if any, net carbon dioxide to the atmosphere and because they greatly reduce ozone formation and carbon monoxide emissions as compared to the environmental output of conventional fuels. (P Bergeron. Environmental impacts of bioethanol. In CE Wyman, ed. Handbook on Bioethanol: Production and Utilization. Washington, DC: Taylor & Francis, 1996, pp 90-103).

Plant biomass is the most abundant source of carbohydrate in the world due to the lignocellulosic materials composing the cell walls of all higher plants. Plant cell walls are divided into two sections, the primary and the secondary cell walls. The primary cell wall, which provides structure for expanding cells (and hence changes as the cell grows), is composed of three major polysaccharides

and one group of glycoproteins. The predominant polysaccharide, and most abundant source of carbohydrates, is cellulose, while hemicellulose and pectin are also found in abundance. Cellulose is a linear beta-(1,4)-D-glucan and comprises 20% to 30% of the primary cell wall by weight. The secondary cell wall, which is produced after the cell has completed growing, also contains polysaccharides and is strengthened through polymeric lignin covalently cross-linked to hemicellulose.

Carbohydrates, and cellulose in particular can be converted to sugars by well-known methods including acid and enzymatic hydrolysis. Enzymatic hydrolysis of cellulose requires the processing of biomass to reduce size and facilitate subsequent handling. Mild acid treatment is then used to hydrolyze part or all of the hemicellulose content of the feedstock. Finally, cellulose is converted to ethanol through the concerted action of cellulases and saccharolytic fermentation (simultaneous saccharification fermentation (SSF)). The SSF process, using the yeast *Saccharomyces cerevisiae* for example, is often incomplete, as it does not utilize the entire sugar content of the plant biomass, namely the hemicellulose fraction.

The cost of producing ethanol from biomass can be divided into three areas of expenditure: pretreatment costs, fermentation costs, and other costs. Pretreatment costs include biomass milling, pretreatment reagents, equipment maintenance, power and water, and waste neutralization and disposal. The fermentation costs can include enzymes, nutrient supplements, yeast, maintenance and scale-up, and waste disposal. Other costs include biomass purchase, transportation and storage, plant labor, plant utilities, ethanol distillation, and administration (which may include technology-use licenses). One of the major expenses incurred in SSF is the cost of the enzymes, as about one kilogram of cellulase is required to fully digest 50 kilograms of cellulose. Economical production of cellulase is also compounded by factors such as the relatively slow gowth rates of cellulase-producing organisms, levels of cellulase expression, and the tendency of enzyme-dependent processes to partially or completely inactivate enzymes due to conditions such as elevated temperature, acidity, proteolytic degradation, and solvent degradation.

Enzymatic degradation of cellulose requires the coordinate action of at least three different types of cellulases. Such enzymes are given an Enzyme Commission (EC) designation according to the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (Eur. J.

Biochem. 264: 607-609 and 610-650, 1999). Endo- beta-(1,4)-glucanases (EC 3.2.1.4) cleave the cellulose strand randomly along its length, thus generating new chain ends. Exo- beta-(1,4)-glucanases (EC 3.2.1.91) are processive enzymes and cleave cellobiosyl units (beta-(1,4)-glucose dimers) from free ends of cellulose strands. Lastly, beta-D-glucosidases (cellobiases: EC 3.2.1.21) hydrolyze cellobiose to glucose. All three of these general activities are required for efficient and complete hydrolysis of a polymer such as cellulose to a subunit, such as the simple sugar, glucose.

Highly thermostable enzymes have been isolated from the cellulolytic thermophile *Acidothermus cellulolyticus gen. nov., sp. nov.*, a bacterium originally isolated from decaying wood in an acidic, thermal pool at Yellowstone National Park. A. Mohagheghi et al., (1986) Int. J. Systematic Bacteriology, 36(3): 435-443. One cellulase enzyme produced by this organism, the endoglucanase EI, is known to display maximal activity at 75 °C to 83°C. M.P. Tucker et al. (1989), Bio/Technology, 7(8): 817-820. E1 endoglucanase has been described in U.S. Patent 5,275,944. The *A. cellulolyticus* E1 endoglucanase is an active cellulase; in combination with the exocellulase CBH I from *Trichoderma reesei,* E1 gives a high level of saccharification and contributes to a degree of synergism. Baker JO et al. (1994), Appl. Biochem. Biotechnol., 45/46: 245-256. The gene coding EI catalytic and carbohydrate binding domains and linker peptide were described in U.S. Patent 5,536,655. E1 has also been expressed as a stable, active enzyme from a wide variety of hosts, including *E. coli*, *Streptomyces lividans*, *Pichia pastoris*, cotton, tobacco, and *Arabidopsis* (Dai Z, Hooker BS, Anderson DB, Thomas SR. Transgenic Res. 2000 Feb; 9(1):43-54).

The potential exists for the successful, commercial-scale expression of heterologous cellulases, and in particular novel cellulases with or without any one or more desirable properties such as thermal tolerance and resistance to acid inactivation, proteolytic inactivation, and solvent inactivation. Such expression can occur in filamentous fungi, bacteria, and other hosts.

There is a need within the art to generate alternative cellulase enzymes capable of assisting in the commercial-scale processing of cellulose to sugar for use in biofuel production. Against this backdrop the present invention has been developed. The potential exists for the successful, commercial-scale expression of heterologous cellulase polypeptides, and in particular novel cellulase polypeptides with or without any one or more desirable properties such as thermal tolerance, and partial or complete resistance to extreme pH inactivation, proteolytic inactivation, solvent

inactivation, chaotropic agent inactivation, oxidizing agent inactivation, and detergent inactivation. Such expression can occur in fungi, bacteria, and other hosts.

Summary of the Invention

The present invention provides Gux1, a novel member of the glycoside hydrolase (GH) family of enzymes, and in particular a thermal tolerant glycoside hydrolase useful in the degradation of cellulose. Gux1 polypeptides of the invention include those having an amino acid sequence shown in SEQ ID NO:1, as well as polypeptides having substantial amino acid sequence identity to the amino acid sequence of SEQ ID NO:1 and useful fragments thereof, including, a catalytic domain having significant sequence similarity to the GH48 family, a first carbohydrate binding domain (type II) and a second carbohydrate binding domain (type III).

The invention also provides a polynucleotide molecule encoding Gux1 polypeptides and fragments of Gux1 polypeptides, for example catalytic and carbohydrate binding domains. Polynucleotide molecules of the invention include those molecules having a nucleic acid sequence as shown in SEQ ID NO:2; those that hybridize to the nucleic acid sequence of SEQ ID NO:2 under high stringency conditions; and those having substantial nucleic acid identity with the nucleic acid sequence of SEQ ID NO:2.

The invention includes variants and derivatives of the Gux1 polypeptides, including fusion proteins. For example, fusion proteins of the invention include Gux1 polypeptide fused to a heterologous protein or peptide that confers a desired function. The heterologous protein or peptide can facilitate purification, oligomerization, stabilization, or secretion of the Gux1 polypeptide, for example. As further examples, the heterologous polypeptide can provide enhanced activity, including catalytic or binding activity, for Gux1 polypeptides, where the enhancement is either additive or synergistic. A fusion protein of an embodiment of the invention can be produced, for example, from an expression construct containing a polynucleotide molecule encoding Gux1 polypeptide in frame with a polynucleotide molecule for the heterologous protein. Embodiments of the invention also comprise vectors, plasmids, expression systems, host cells, and the like, containing a Gux1 polynucleotide molecule. Genetic engineering methods for the production of Gux1 polypeptides of embodiments of the invention include expression of a polynucleotide molecule in cell free expression systems and in cellular hosts, according to known methods.

The invention further includes compositions containing a substantially purified Gux1 polypeptide of the invention and a carrier. Such compositions are administered to a biomass containing cellulose for the reduction or degradation of the cellulose.

The invention also provides reagents, compositions, and methods that are useful for analysis of Gux1 activity.

These and various other features as well as advantages which characterize the present invention will be apparent from a reading of the following detailed description and a review of the associated drawings.

The following Tables 4 and 5 includes sequences used in describing embodiments of the present invention. In Table 4, the abbreviations are as follows: CD, catalytic domain; CBD_II, carbohydrate binding domain type II; CBD_III, carbohydrate binding domain type III; and FN-III, fibronectin domain type III. When used herein, N* indicates a string of unknown nucleic acid units, and X* indicates a string of unknown amino acid units, for example about 50 or more. Table 4 includes approximate start and stop information for segments, and Table 5 includes amino acid sequence data for segments.

Table 4. Nucleotide and polypeptide segments.

Gux1 Segment	base BEGIN	base END	Length, bp	aa BEGIN No.	aa	aa END No.	aa	Length, aa	SEQ ID No. (amino acid)	SEQ ID No. (nucleotide)
Total length	1	3366	3366	1	M	1121	S	1121	1	2
Signal (potential)	1	102	102	1	M	34	A	34	3	
CBD III	103	561	459	35	A	187	A	153	4	
CD (GH48)	691	2610	1920	231	N	870	P	640	5	
FN-III	2701	2955	255	901	D	985	G	85	6	
CBD II	3061	3363	303	1021	G	1121	S	101	7	

Table 5. Gene/polypeptide segments with amino acid sequences.

SEQ ID No. (amino acid)	SEQ ID No. (nucleotide)	Gux1 Segment	segment data
1	2	Total length	SEQ ID NO: 1 (see Table 1); SEQ ID NO: 2 (see Table 2)
3		Signal (potential)	MPGLRRRLRAGIVSAAALGSLVSGLVAVAPVAHA
4		CBD_III	AVTLKAQYKNNDSAPSDNQIKPGLQLVNTGSSSVDLSTVTVRYWFTRDGGSSTLVYNCDWAA MGCGNIRASFGSVNPATPTADTYLQLSFTGGTLAAGGSTGEIQNRVNKSDWSNFDETNDYSY GTNTTFQDWTKVTVYVNGVLVWGTEPSGA
5		CD (GH48)	NDPYIQRFLTMYNKIHDPANGYFSPQGIPYHSVETLIVEAPDYGHETTSEAYSFWLWLEATY GAVTGNWTPFNNAWTTMETYMIPQHADQPNNASYNPNSPASYAPEEPLPSMYPVAIDSSVPV GHDPLAAELQSTYGTPDIYGMHWLADVDNIYGYGDSPGGGCELGPSAKGVSYINTFQRGSQE SVWETVTQPTCDNGKYGGAHGYVDLFIQGSTPPQWKYTDAPDADARAVQAAYWAYTWASAQG KASAIAPTIAKASQTGDYLRYSLFDKYFKQVGNCYPASSCPGATGRQSETYLIGWYYAWGGS SQGWAWRIGDGAAHFGYQNPLAAWAMSNVTPLIPLSPTAKSDWAASLQRQLEFYQWLQSAEG AIAGGATNSWNGNYGTPPAGDSTFYGMAYDWEPVYHDPPSNNWFGFQAWSMERVAEYYYVTG DPKAKALLDKWVAWVKPNVTTGASWSIPSNLSWSGQPDTWNPSNPGTNANLHVTITSSGQDV GVAAALAKTLEYYAAKSGDTASRDLAKGLLDSMWNNDQDSLGVSTPETRTDYSRFTQVYDPT TGDGLYIPSGWTGTMPNGDQIKPGATFLSIRSWYTKDPQWSKVQAYLNGGPAPTFNYHRFWA ESDFAMANADFGMLFPSGSP
6		FN-III	DTTPPSVPTGLQVTGTTTSSVSLSWTASTDNVGVAHYNVYRNGTLVGQPTATSFTDTGLAAG TSYTYTVAAVDAAGNTSAQSFAG
7		CBD_II	GASCTATYVVNSDWGSGFTTTVTVTNTGTRATSGWTVTWSFAGNQTVTNYWNTAL TQSGKSVTAKNLSYNNVIQPGQSTTFGFNGSYSGTNTAPTLSCTAS

Brief Description of the Drawings

FIG. 1 is a schematic representation of the gene sequence and amino acid segment organization.

FIG 2 is a graphic representation of the glycoside hydrolase gene/protein families found in various organisms.

Detailed Description

Definitions:

The following definitions are provided to facilitate understanding of certain terms used frequently herein and are not meant to limit the scope of the present disclosure:

"Amino acid" refers to any of the twenty naturally occuring amino acids as well as any modified amino acid sequences. Modifications may include natural processes such as posttranslational processing, or may include chemical modifications which are known in the art. Modifications include but are not limited to: phosphorylation, ubiquitination, acetylation, amidation, glycosylation, covalent attachment of flavin, ADP-ribosylation, cross linking, iodination, methylation, and alike.

"Antibody" refers to a Y-shaped molecule having a pair of antigen binding sites, a hinge region and a constant region. Fragments of antibodies, for example an antigen binding fragment (Fab), chimeric antibodies, antibodies having a human constant region coupled to a murine antigen binding region, and fragments thereof, as well as other well known recombinant antibodies are included in the present invention.

"Antisense" refers to polynucleotide sequences that are complementary to target "sense" polynucleotide sequence.

"Binding activity" refers to any activity that can be assayed by characterizing the ability of a polypeptide to bind to a substrate. The substrate can be a polymer such as cellulose or can be a complex molecule or aggregate of molecules where the entire moiety comprises at least some cellulose.

"Cellulase activity" refers to any activity that can be assayed by characterizing the enzymatic activity of a cellulase. For example, cellulase activity can be assayed by determining how much reducing sugar is produced during a fixed amount of time for a set amount of enzyme (see Irwin et al., (1998) *J. Bacteriology*, 1709-1714). Other assays are well known in the art and can be substituted.

"Complementary" or "complementarity" refers to the ability of a polynucleotide in a polynucleotide molecule to form a base pair with another polynucleotide in a second polynucleotide molecule. For example, the sequence A-G-T is complementary to the sequence T-C-A. Complementarity may be partial, in which only some of the polynucleotides match according to base pairing, or complete, where all the polynucleotides match according to base pairing.

"Expression" refers to transcription and translation occurring within a host cell. The level of expression of a DNA molecule in a host cell may be determined on the basis of either the amount of corresponding mRNA that is present within the cell or the amount of DNA molecule encoded protein produced by the host cell (Sambrook et al., 1989, *Molecular cloning: A Laboratory Manual*, 18.1-18.88).

"Fusion protein" refers to a first protein having attached a second, heterologous protein. Preferably, the heterologous protein is fused via recombinant DNA techniques, such that the first and second proteins are expressed in frame. The heterologous protein can confer a desired characteristic to the fusion protein, for example, a detection signal, enhanced stability or stabilization of the protein, facilitated oligomerization of the protein, or facilitated purification of the fusion protein. Examples of heterologous proteins useful in the fusion proteins of the invention include molecules having one or more catalytic domains of Gux1, one or more binding domains of Gux1, one or more catalytic domains of a glycoside hydrolase other than Gux1, one or more binding domains of a glycoside hydrolase other than Gux1, or any combination thereof. Further examples include immunoglobulin molecules and portions thereof, peptide tags such as histidine tag (6-His), leucine zipper, substrate targeting moieties, signal peptides, and the like. Fusion proteins are also meant to encompass variants and derivatives of Gux1 polypeptides that are generated by conventional site-directed mutagenesis and more modern techniques such as directed evolution, discussed infra.

"Genetically engineered" refers to any recombinant DNA or RNA method used to create a prokaryotic or eukaryotic host cell that expresses a protein at elevated levels, at lowered levels, or in a mutated form. In other words, the host cell has been transfected, transformed, or transduced with a recombinant polynucleotide molecule, and thereby been altered so as to cause the cell to alter expression of the desired protein. Methods and vectors for genetically engineering host cells are well known in the art; for example various techniques are illustrated in Current Protocols in Molecular Biology, Ausubel et al., eds. (Wiley & Sons, New York, 1988, and quarterly updates). Genetically engineering techniques include but are not limited to expression vectors, targeted homologous recombination and gene activation (see, for example, U.S. Patent No. 5,272,071 to Chappel) and trans activation by engineered transcription factors (see, for example, Segal et al., 1999, *Proc Natl Acad Sci USA* 96(6):2758-63).

"Glycoside hydrolase family" refers to a family of enzymes which hydrolyze the glycosidic bond between two or more carbohydrates or between a carbohydrate and a non-carbohydrate moiety (Henrissat B., (1991) Biochem. J., 280:309-316). Identification of a putative glycoside hydrolase family member is made based on an amino acid sequence comparison and the finding of significant sequence similarity within the putative member's catalytic domain, as compared to the catalytic domains of known family members.

"Homology" refers to a degree of complementarity between polynucleotides, having significant effect on the efficiency and strength of hybridization between polynucleotide molecules. The term also can refer to a degree of similarity between polypeptides. Two polypeptides having greater than or equal to about 60% similarity are presumptively homologous.

"Host cell" or "host cells" refers to cells expressing a heterologous polynucleotide molecule. Host cells of the present invention express polynucleotides encoding Gux1 or a fragment thereof. Examples of suitable host cells useful in the present invention include, but are not limited to, prokaryotic and eukaryotic cells. Specific examples of such cells include bacteria of the genera *Escherichia, Bacillus,* and *Salmonella,* as well as members of the genera *Pseudomonas, Streptomyces,* and *Staphylococcus*; fungi, particularly filamentous fungi such as *Trichoderma* and *Aspergillus*, *Phanerochaete chrysosporium* and other white rot fungi; also other fungi including *Fusaria*, molds, and yeast including *Saccharomyces* sp., *Pichia* sp., and *Candida* sp. and the like; plants e.g.

Arabidopsis, cotton, barley, tobacco, potato, and aquatic plants and the like; SF9 insect cells (Summers and Smith, 1987, *Texas Agriculture Experiment Station Bulletin*, 1555), and the like. Other specific examples include mammalian cells such as human embyonic kidney cells (293 cells), Chinese hamster ovary (CHO) cells (Puck et al., 1958, *Proc. Natl. Acad. Sci. USA* 60, 1275-1281), human cervical carcinoma cells (HELA) (ATCC CCL 2), human liver cells (Hep G2) (ATCC HB8065), human breast cancer cells (MCF-7) (ATCC HTB22), human colon carcinoma cells (DLD-1) (ATCC CCL 221), Daudi cells (ATCC CRL-213), murine myeloma cells such as P3/NSI/1-Ag4-1 (ATCC TIB-18), P3X63Ag8 (ATCC TIB-9), SP2/0-Ag14 (ATCC CRL-1581) and the like.

"Hybridization" refers to the pairing of complementary polynucleotides during an annealing period. The strength of hybridization between two polynucleotide molecules is impacted by the homology between the two molecules, stringency of the conditions involved, the melting temperature of the formed hybrid and the G:C ratio within the polynucleotides.

"Identity" refers to a comparison between pairs of nucleic acid or amino acid molecules. Methods for determining sequence identity are known. See, for example, computer programs commonly employed for this purpose, such as the Gap program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University Research Park, Madison Wisconsin), that uses the algorithm of Smith and Waterman, 1981, *Adv. Appl. Math., 2:* 482-489.

"Isolated" refers to a polynucleotide or polypeptide that has been separated from at least one contaminant (polynucleotide or polypeptide) with which it is normally associated. For example, an isolated polynucleotide or polypeptide is in a context or in a form that is different from that in which it is found in nature.

"Nucleic acid sequence" refers to the order or sequence of deoxyribonucleotides along a strand of deoxyribonucleic acid. The order of these deoxyribonucleotides determines the order of amino acids along a polypeptide chain. The deoxyribonucleotide sequence thus codes for the amino acid sequence.

"Polynucleotide" refers to a linear sequence of nucleotides. The nucleotides may be ribonucleotides, or deoxyribonucleotides, or a mixture of both. Examples of polynucleotides in the context of the

present invention include single and double stranded DNA, single and double stranded RNA, and hybrid molecules having mixtures of single and double stranded DNA and RNA. The polynucleotides of the present invention may contain one or more modified nucleotides.

"Protein," "peptide," and "polypeptide" are used interchangeably to denote an amino acid polymer or a set of two or more interacting or bound amino acid polymers.

"Purify," or "purified" refers to a target protein that is free from at least 5-10% of contaminating proteins. Purification of a protein from contaminating proteins can be accomplished using known techniques, including ammonium sulfate or ethanol precipitation, acid precipitation, heat precipitation, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography, size-exclusion chromatography, and lectin chromatography. Various protein purification techniques are illustrated in *Current Protocols in Molecular Biology*, Ausubel et al., eds. (Wiley & Sons, New York, 1988, and quarterly updates).

"Selectable marker" refers to a marker that identifies a cell as having undergone a recombinant DNA or RNA event. Selectable markers include, for example, genes that encode antimetabolite resistance such as the DHFR protein that confers resistance to methotrexate (Wigler et al, 1980, *Proc Natl Acad Sci* USA 77:3567; O'Hare et al., 1981, *Proc Natl Acad Sci USA*, 78:1527), the GPT protein that confers resistance to mycophenolic acid (Mulligan & Berg, 1981, *PNAS USA*, 78:2072), the neomycin resistance marker that confers resistance to the aminoglycoside G-418 (Calberre-Garapin et al., 1981, *J Mol Biol*, 150:1), the Hygro protein that confers resistance to hygromycin (Santerre et al., 1984, *Gene* 30:147), and the Zeocin™ resistance marker (Invitrogen). In addition, the herpes simplex virus thymidine kinase, hypoxanthine-guanine phosphoribosyltransferase and adenine phosphoribosyltransferase genes can be employed in tk^-, $hgprt^-$ and $aprt^-$ cells, respectively.

"Stringency" refers to the conditions (temperature, ionic strength, solvents, etc) under which hybridization between polynucleotides occurs. A hybridzation reaction conducted under high stringency conditions is one that will only occur between polynucleotide molecules that have a high degree of complementary base pairing (85% to 100% identity). Conditions for high stringency hybridization, for example, may include an overnight incubation at about 42°C for about 2.5 hours in 6

X SSC/0.1% SDS, followed by washing of the filters in 1.0 X SSC at 65°C, 0.1% SDS. A hybridization reaction conducted under moderate stringency conditions is one that will occur between polynucleotide molecules that have an intermediate degree of complementary base pairing (50% to 84% identity).

"Substrate targeting moiety" refers to any signal on a substrate, either naturally occurring or genetically engineered, used to target any Gux1 polypeptide or fragment thereof to a substrate. Such targeting moieties include ligands that bind to a substrate structure. Examples of ligand/receptor pairs include carbohydrate binding domains and cellulose. Many such substrate-specific ligands are known and are useful in the present invention to target a Gux1 polypeptide or fragment thereof to a substrate. A novel example is a Gux1 carbohydrate binding domain that is used to tether other molecules to a cellulose-containing substrate such as a fabric.

"Thermal tolerant" refers to the property of withstanding partial or complete inactivation by heat and can also be described as thermal resistance or thermal stability. Although some variation exists in the literature, the following definitions can be considered typical for the optimum temperature range of stability and activity for enzymes: psycrophilic (below freezing to 10C); mesophilic (10°C to 50°C); thermophilic (50°C to 75°C); and caldophilic (75°C to above boiling water temperature). The stability and catalytic activity of enzymes are linked characteristics, and the ways of measuring these properties vary considerably. For industrial enzymes, stability and activity are best measured under use conditions, often in the presence of substrate. Therefore, cellulases that must act on process streams of cellulose must be able to withstand exposure up to thermophilic or even caldophilic temperatures for digestion times in excess of several hours.

In encompassing a wide variety of potential applications for embodiments of the present invention, thermal tolerance refers to the ability to function in a temperature range of from about 15°C to about 100°C. A preferred range is from about 30°C to about 80°C. A highly preferred range is from about 50°C to about 70°C. For example, a protein that can function at about 45°C is considered in the preferred range even though it may be susceptible to partial or complete inactivation at temperatures in a range above about 45°C and less than about 80°C. For polypeptides derived from organisms such as *Acidothermus*, the desirable property of thermal tolerance among is often accompanied by

other desirable characteristics such as: resistance to extreme pH degradation, resistance to solvent degradation, resistance to proteolytic degradation, resistance to detergent degradation, resistance to oxidizing agent degradation, resistance to chaotropic agent degradation, and resistance to general degradation. Cowan DA in Danson MJ et al. (1992) The Archaebacteria, Biochemistry and Biotechnology at 149-159, University Press, Cambridge, ISBN 1855780100. Here 'resistance' is intended to include any partial or complete level of residual activity. When a polypeptide is described as thermal tolerant it is understood that any one, more than one, or none of these other desirable properties can be present.

"Variant", as used herein, means a polynucleotide or polypeptide molecule that differs from a reference molecule. Variants can include nucleotide changes that result in amino acid substitutions, deletions, fusions, or truncations in the resulting variant polypeptide when compared to the reference polypeptide.

"Vector," "extra-chromosomal vector" or "expression vector" refers to a first polynucleotide molecule, usually double-stranded, which may have inserted into it a second polynucleotide molecule, for example a foreign or heterologous polynucleotide. The heterologous polynucleotide molecule may or may not be naturally found in the host cell, and may be, for example, one or more additional copy of the heterologous polynucleotide naturally present in the host genome. The vector is adapted for transporting the foreign polynucleotide molecule into a suitable host cell. Once in the host cell, the vector may be capable of integrating into the host cell chromosomes. The vector may optionally contain additional elements for selecting cells containing the integrated polynucleotide molecule as well as elements to promote transcription of mRNA from transfected DNA. Examples of vectors useful in the methods of the present invention include, but are not limited to, plasmids, bacteriophages, cosmids, retroviruses, and artificial chromosomes.

Within the application, unless otherwise stated, the techniques utilized may be found in any of several well-known references, such as: *Molecular Cloning: A Laboratory Manual* (Sambrook et al. (1989) Molecular cloning: A Laboratory Manual), *Gene Expression Technology* (Methods in Enzymology, Vol. 185, edited by D. Goeddel, 1991 Academic Press, San Diego, CA), "Guide to Protein Purification" in *Methods in Enzymology* (M.P. Deutshcer, 3d., (1990) Academic Press, Inc.), *PCR Protocols: A Guide to Methods and Applications* (Innis et al. (1990) Academic Press, San

Diego, CA), Culture of Animal Cells: A Manual of Basic Technique, 2nd ed. (R.I. Freshney (1987) Liss, Inc., New York, NY), *and Gene Transfer and Expression Protocols*, pp 109-128, ed. E.J. Murray, The Humana Press Inc., Clifton, N.J.).

O-Glycoside Hydrolases:

Glycoside hydrolases are a large and diverse family of enzymes that hydrolyse the glycosidic bond between two carbohydrate moieties or between a carbohydrate and a non-carbohydrate moiety (See FIG. 2). Glycoside hydrolase enzymes are classified into glycoside hydrolase (GH) families based on significant amino acid similarities within their catalytic domains. Enzymes having related catalytic domains are grouped together within a family, (Henrissat et al., (1991) supra, and Henrissat et al. (1996), Biochem. J. 316:695-696), where the underlying classification provides a direct relationship between the GH domain amino acid sequence and how a GH domain will fold. This information ultimately provides a common mechanism for how the enzyme will hydrolyse the glycosidic bond within a substrate, *i.e.*, either by a retaining mechanism or inverting mechanism (Henrissat., B, (1991) supra).

Cellulases belong to the GH family of enzymes. Cellulases are produced by a variety of bacteria and fungi to degrade the β-1,4 glycosidic bond of cellulose and to so produce successively smaller fragments of cellulose and ultimately produce glucose. At present, cellulases are found within are at least 11 different GH families. Three different types of cellulase enzyme activities have been identified within these GH families: exo-acting cellulases which cleave successive disaccharide units from the non-reducing ends of a cellulose chain; endo-acting cellulases which randomly cleave successive disaccharide units within the cellulose chain; and β-glucosidases which cleave successive disaccharide units to glucose (J. W. Deacon, (1997) Modern Mycology, 3rd Ed., ISBN: 0-632-03077-1, 97-98).

Many cellulases are characterized by having a multiple domain unit within their overall structure, a GH or catalytic domain is joined to a carbohydrate-binding domain (CBD) by a glycosylated linker peptide (see FIG. 1) (Koivula et al., (1996) Protein Expression and Purification 8:391-400). As noted above, cellulases do not belong to any one family of GH domains, but rather have been identified within at least 11 different GH families to date. The CBD type domain increases the

concentration of the enzyme on the substrate, in this case cellulose, and the linker peptide provides flexibility for both larger domains.

Conversion of cellulose to glucose is an essential step in the production of ethanol or other biofuels from biomass. Cellulases are an important component of this process, where approximately one kilogram of cellulase can digest fifty kilograms of cellulose. Within this process, thermostable cellulases have taken precedent, due to their ability to function at elevated temperatures and under other conditions including pH extremes, solvent presence, detergent presence, proteolysis, etc. (see Cowan DA (1992), supra).

Highly thermostable cellulase enzymes are secreted by the cellulolytic themophile *Acidothermus cellulolyticus* (U.S. Patent Nos. 5,275,944 and 5,110,735). This bacterium was originally isolated from decaying wood in an acidic, thermal pool at Yellowstone National Park and deposited with the American Type Culture Collection (ATCC 43068) (Mohagheghi et al., (1986) *Int. J. System. Bacteriol.*, 36:435-443).

Recently, a thermostable cellulase, E1 endoglucanase, was identified and characterized from *Acidothermus cellulolyticus* (U.S. Patent No. 5,536,655). The E1 endoglucanase has maximal activity between 75 and 83°C and is active to a pH well below 5. Thermostable cellulase, and E1 endoglucanase, are useful in the conversion of biomass to biofuels, and in particular, are useful in the conversion of cellulose to glucose. Conversion of biomass to biofuel represents an extremely important alternative fuel source that is more environmentally friendly than conventional fuels, and provides a use, in some cases, for waste products.

Gux1:

As described more fully in the Examples below, Gux1, a novel thermostable cellulase, has now been identified and characterized. The predicted amino acid sequence of Gux1 (SEQ ID NO:1) has an organization characteristic of a cellulase enzyme. Gux1 contains a carbohydrate binding domain - linker domain - catalytic domain -linker domain- fibronectin domain - linker domain - carbohydrate binding domain unit. In particular, a catalytic domain unit includes a carbohydrate binding domain type III (amino acids from about A35 to about A187), a GH48 catalytic domain (amino acids from about N231 to about P870), and a CBD_{II} (amino acids from about G1021 to about S1121). As

discussed in more detail below, significant amino acid similarity of Gux1 to other cellulases identifies Gux1 as a cellulase.

Gux1, as noted above, has a catalytic domain, identified as belonging to the GH48 family. The GH48 domain family includes a number of exoglucanases, for example, from *Cellulomonas fimi*, and exoglucanase E3 isolated from *Thermobifida fusca*. The GH48 members degrade substrate using an inverting mechanism. Being a member of the GH48 family of proteins identifies Gux1 as potentially having exoglucanase activity. In addition, the predicted amino acid sequence (SEQ ID NO: 1) indicates that CBD type II and CBD type III domains are present as characterized by Tomme P. et al. (1995), in Enzymatic Degradation of Insoluble Polysaccharides (Saddler JN & Penner M, eds.), at 142-163, American Chemical Society, Washington. See also Tomme, P. & Claeyssens, M. (1989) FEBS Lett. 243, 239-2431; Gilkes, N.R et al., (1988) J.Biol.Chem. 263, 10401-10407.

Gux1 is also a thermostable cellulase as it is produced by the themophile *Acidothermus cellulolyticus*. As discussed, Gux1 polypeptides can have other desirable characteristics (see Cowan DA (1992), supra). Like other members of the cellulase family, and in particular thermostable cellulases, Gux1 polypeptides are useful in the conversion of biomass to biofuels and biofuel additives, and in particular, biofuels from cellulose. It is envisioned that Gux1 polypeptides could be used for other purposes, for example in detergents, pulp and paper processing, food and feed processing, and in textile processes. Gux1 polypeptides can be used alone or in combination with one or more other cellulases or glycoside hydrolases to perform the uses described herein or known within the relevant art, all of which are within the scope of the present disclosure.

Gux1 Polypeptides:

Gux1 polypeptides of the invention include isolated polypeptides having an amino acid sequence as shown below in Example 1; Table 1 and in SEQ ID NO:1, as well as variants and derivatives, including fragments, having substantial identity to the amino acid sequence of SEQ ID NO:1 and that retain any of the functional activities of Gux1. Gux1 polypeptide activity can be determined, for example, by subjecting the variant, derivative, or fragment to a substrate binding assay or a cellulase activity assay such as those described in Irwin D et al., J. Bacteriology 180(7): 1709-1714 (April 1998).

Table 1. Gux1 amino acid sequence (SEQ ID NO:1)

```
MPGLRRRLRAGIVSAAALGSLVSGLVAVAPVAHAAVTLKAQYKNNDSAPS
DNQIKPGLQLVNTGSSSVDLSTVTVRYWFTRDGGSSTLVYNCDWAAMGCG
NIRASFGSVNPATPTADTYLQLSFTGGTLAAGGSTGEIQNRVNKSDWSNF
DETNDYSYGTNTTFQDWTKVTVYVNGVLVWGTEPSGATASPSASATPSPS
SSPTTSPSSSPSPSSSPTPTPSSSSPPPSSNDPYIQRFLTMYNKIHDPAN
GYFSPQGIPYHSVETLIVEAPDYGHETTSEAYSFWLWLEATYGAVTGNWT
PFNNAWTTMETYMIPQHADQPNNASYNPNSPASYAPEEPLPSMYPVAIDS
SVPVGHDPLAAELQSTYGTPDIYGMHWLADVDNIYGYGDSPGGGCELGPS
AKGVSYINTFQRGSQESVWETVTQPTCDNGKYGGAHGYVDLFIQGSTPPQ
WKYTDAPDADARAVQAAYWAYTWASAQGKASAIAPTIAKASQTGDYLRYS
LFDKYFKQVGNCYPASSCPGATGRQSETYLIGWYYAWGGSSQGWAWRIGD
GAAHFGYQNPLAAWAMSNVTPLIPLSPTAKSDWAASLQRQLEFYQWLQSA
EGAIAGGATNSWNGNYGTPPAGDSTFYGMAYDWEPVYHDPPSNNWFGFQA
WSMERVAEYYYVTGDPKAKALLDKWVAWVKPNVTTGASWSIPSNLSWSGQ
PDTWNPSNPGTNANLHVTITSSGQDVGVAAALAKTLEYYAAKSGDTASRD
LAKGLLDSMWNNDQDSLGVSTPETRTDYSRFTQVYDPTTGDGLYIPSGWT
GTMPNGDQIKPGATFLSIRSWYTKDPQWSKVQAYLNGGPAPTFNYHRFWA
ESDFAMANADFGMLFPSGSPSPTPSPTPTSSPSPTPSSSPTPSPSPSPTG
DTTPPSVPTGLQVTGTTTSSVSLSWTASTDNVGVAHYNVYRNGTLVGQPT
ATSFTDTGLAAGTSYTYTVAAVDAAGNTSAQSFAGDSDDGIAVASPSPSP
TPTSSPSPTPSPTPSPTSTSGASCTATYVVNSDWGSGFTTTVTVTNTGTR
ATSGWTVTWSFAGNQTVTNYWNTALTQSGKSVTAKNLSYNNVIQPGQSTT
FGFNGSYSGTNTAPTLSCTAS
```

As listed and described in Tables 1 and 5, the isolated Gux1 polypeptide includes an N-terminal hydrophobic region that functions as a signal peptide, having an amino acid sequence that begins with Met1 and extends to about A34; a carbohydrate binding domain having sequence similarity to such type III domains that begins with about A35 and extends to about A187, a catalytic domain having significant sequence similarity to a GH48 family domain that begins with about N231 and extends to about P870, a fibronectin type III domain that begins with about D901 and extends to about G985, a carbohydrate binding domain type II region that begins with about G1021 and extends to about S1121. Variants and derivatives of Gux1 include, for example, Gux1 polypeptides modified by covalent or aggregative conjugation with other chemical moieties, such as glycosyl groups, polyethylene glycol (PEG) groups, lipids, phosphate, acetyl groups, and the like.

The amino acid sequence of Gux1 polypeptides of the invention is preferably at least about 60% identical, more preferably at least about 70% identical, or in some embodiments at least about 90% identical, to the Gux1 amino acid sequence shown above in Table 1 and SEQ ID NO:1. The percentage identity, also termed homology (see definition above) can be readily determined, for example, by comparing the two polypeptide sequences using any of the computer programs commonly employed for this purpose, such as the Gap program (Wisconsin Sequence Analysis Package, Version 8 for Unix,

Genetics Computer Group, University Research Park, Madison Wisconsin), which uses the algorithm of Smith and Waterman, 1981, *Adv. Appl. Math.* 2: 482-489.

Variants and derivatives of the Gux1 polypeptide may further include, for example, fusion proteins formed of a Gux1 polypeptide and a heterologous polypeptide. Preferred heterologous polypeptides include those that facilitate purification, oligomerization, stability, or secretion of the Gux1 polypeptides.

Gux1 polypeptide fragments may include, but are not limited to, the polypeptide sequences listed in Table 4, SEQ ID NOS: 3, 4, 5, 6, and 7.

Gux1 polypeptide variants and derivatives, as used in the description of the invention, can contain conservatively substituted amino acids, meaning that one or more amino acid can be replaced by an amino acid that does not alter the secondary and/or tertiary structure of the polypeptide. Such substitutions can include the replacement of an amino acid, by a residue having similar physicochemical properties, such as substituting one aliphatic residue (Ile, Val, Leu, or Ala) for another, or substitutions between basic residues Lys and Arg, acidic residues Glu and Asp, amide residues Gln and Asn, hydroxyl residues Ser and Tyr, or aromatic residues Phe and Tyr. Phenotypically silent amino acid exchanges are described more fully in Bowie *et al.,* 1990, *Science 247*:1306-1310. In addition, functional Gux1 polypeptide variants include those having amino acid substitutions, deletions, or additions to the amino acid sequence outside functional regions of the protein, for example, outside the catalytic and carbohydrate binding domains. These would include, for example, the various linker sequences that connect functional domains as defined herein.

The Gux1 polypeptides of the present invention are preferably provided in an isolated form, and preferably are substantially purified. The polypeptides may be recovered and purified from recombinant cell cultures by known methods, including, for example, ammonium sulfate or ethanol precipitation, anion or cation exchange chromatography, phosphocellulose chromatography, hydrophobic interaction chromatography, affinity chromatography, hydroxylapatite chromatography, and lectin chromatography. Preferably, high performance liquid chromatography (HPLC) is employed for purification.

Another preferred form of Gux1 polypeptides is that of recombinant polypeptides as expressed by suitable hosts. Furthermore, the hosts can simultaneously produce other cellulases such that a mixture is produced comprising a Gux1 polypeptide and one or more other cellulases. Such a mixture can be effective in crude fermentation processing or other industrial processing.

Gux1 polypeptides can be fused to heterologous polypeptides to facilitate purification. Many available heterologous peptides (peptide tags) allow selective binding of the fusion protein to a binding partner. Non-limiting examples of peptide tags include 6-His, thioredoxin, hemaglutinin, GST, and the OmpA signal sequence tag. A binding partner that recognizes and binds to the heterologous peptide can be any molecule or compound, including metal ions (for example, metal affinity columns), antibodies, antibody fragments, or any protein or peptide that preferentially binds the heterologous peptide to permit purification of the fusion protein.

Gux1 polypeptides can be modified to facilitate formation of Gux1 oligomers. For example, Gux1 polypeptides can be fused to peptide moieties that promote oligomerization, such as leucine zippers and certain antibody fragment polypeptides, for example, Fc polypeptides. Techniques for preparing these fusion proteins are known, and are described, for example, in WO 99/31241 and in Cosman et.al., 2001 *Immunity* 14:123-133. Fusion to an Fc polypeptide offers the additional advantage of facilitating purification by affinity chromatography over Protein A or Protein G columns. Fusion to a leucine-zipper (LZ), for example, a repetitive heptad repeat, often with four or five leucine residues interspersed with other amino acids, is described in Landschultz et al., 1988, *Science*, 240:1759.

It is also envisioned that an expanded set of variants and derivatives of Gux1 polynucleotides and/or polypeptides can be generated to select for useful molecules, where such expansion is achieved not only by conventional methods such as site-directed mutagenesis (SDM) but also by more modern techniques, either independently or in combination.

Site-directed-mutagenesis is considered an informational approach to protein engineering and can rely on high-resolution crystallographic structures of target proteins and some stratagem for specific amino acid changes (Van Den Burg, B.; Vriend, G.; Veltman, O.R.; Venema, G.; Eijsink, V.G.H. Proc. Nat. Acad. Sci. U.S. 1998, 95, 2056-2060). For example, modification of the amino acid sequence of Gux1 polypeptides can be accomplished as is known in the art, such as by introducing mutations at particular

locations by oligonucleotide-directed mutagenesis (Walder et al.,1986, Gene, 42:133; Bauer et al., 1985, Gene 37:73; Craik, 1985, BioTechniques, 12-19; Smith et al., 1981, Genetic Engineering: Principles and Methods, Plenum Press; and U.S. Patent No. 4,518,584 and U.S. Patent No. 4,737,462). SDM technology can also employ the recent advent of computational methods for identifying site-specific changes for a variety of protein engineering objectives (Hellinga, H.W. Nature Structural. Biol. 1998, 5, 525-527).

The more modern techniques include, but are not limited to, non-informational mutagenesis techniques (referred to generically as "directed evolution"). Directed evolution, in conjunction with high-throughput screening, allows testing of statistically meaningful variations in protein conformation (Arnold, F.H. Nature Biotechnol. 1998, 16, 617-618). Directed evolution technology can include diversification methods similar to that described by Crameri A. et al. (1998, Nature 391: 288-291), site-saturation mutagenesis, staggered extension process (StEP) (Zhao, H.; Giver, L.; Shao, Z.; Affholter, J.A.; Arnold, F.H. Nature Biotechnol. 1998, 16, 258-262), and DNA synthesis/reassembly (U.S. Patent 5,965,408).

Fragments of the Gux1 polypeptide can be used, for example, to generate specific anti-Gux1 antibodies. Using known selection techniques, specific epitopes can be selected and used to generate monoclonal or polyclonal antibodies. Such antibodies have utlilty in the assay of Gux1 activity as well as in purifying recombinant Gux1 polypeptides from genetically engineered host cells.

Gux1 Polynucleotides:

The invention also provides polynucleotide molecules encoding the Gux1 polypeptides discussed above. Gux1 polynucleotide molecules of the invention include polynucleotide molecules having the nucleic acid sequence shown in Table 2 and SEQ ID NO:2, polynucleotide molecules that hybridize to the nucleic acid sequence of Table 2 and SEQ ID NO:2 under high stringency hybridization conditions (for example, 42°, 2.5 hr., 6X SCC, 0.1%SDS); and polynucleotide molecules having substantial nucleic acid sequence identity with the nucleic acid sequence of Table 2 and SEQ ID NO:2, particularly with those nucleic acids encoding the catalytic domain, GH48 (from about amino acid N231 to about P870), the carbohydrate binding domain III (from about amino acid A35 to A187) and carbohydrate binding domain II (from about G1021 to about amino acid S1121).

Table 2. Gux1 nucleotide sequence (SEQ ID NO:2).

ATGCCAGGATTACGACGGCGACTCCGCGCCGGTATCGTCTCGGCGGCGGCGTTGGGGTCGCTGGTTAGCGG
GCTCGTTGCCGTCGCACCAGTCGCGCACGCGGCGGTGACTCTCAAAGCGCAGTATAAGAACAATGATTCGG
CGCCGAGTGACAACCAGATCAAACCGGGTCTCCAGTTGGTGAATACCGGGTCGTCGTCGGTGGATTTGTCG
ACGGTGACGGTGCGGTACTGGTTCACCCGGGATGGTGGGTCGTCGACACTGGTGTACAACTGTGACTGGGC
GGCGATGGGGTGTGGGAATATCCGCGCCTCGTTCGGCTCGGTGAACCCGGCGACGCCGACGGCGGACACC
TACCTGCAGTTGTCGTTCACTGGTGGAACGTTGGCCGCTGGTGGGTCGACGGGTGAGATTCAAAACCGGGT
GAATAAGAGTGACTGGTCGAACTTTGATGAGACCAATGACTACTCGTATGGGACGAACACCACCTTCCAGG
ACTGGACGAAGGTGACGGTGTACGTCAACGGCGTGTTGGTCTGGGGGACCGAACCGTCCGGAGCGACGGC
GTCTCCATCCGCGTCGGCGACGCCCAGCCCGTCCAGTTCACCGACCACGAGTCCGAGTTCGTCCCCGTCGCC
GAGCAGCAGCCCGACGCCGACACCGAGCAGCTCGTCGCCGCCCCGTCGTCCAACGACCCGTACATCCAGCG
GTTCCTCACGATGTACAACAAGATTCACGACCCAGCGAACGGCTACTTCAGCCCGCAGGGAATTCCCTACC
ACTCGGTAGAAACGCTCATCGTTGAGGCACCGGACTACGGGCACGAGACAACTTCGGAGGCGTACAGCTTC
TGGCTCTGGCTCGAAGCGACGTACGGCGCAGTGACCGGCAACTGGACGCCGTTCAACAACGCCTGGACGAC
GATGGAAACGTACATGATCCCGCAGCACGCGGACCAGCCGAACAACGCGTCGTACAACCCCAACAGCCCG
GCGTCGTACGCTCCGGAAGAGCCGCTGCCCAGCATGTACCCGGTTGCCATCGACAGCAGCGTGCCGGTTGG
GCACGACCCGCTCGCCGCCGAATTGCAGTCGACGTACGGCACTCCGGACATTTACGGCATGCACTGGCTGG
CCGACGTTGACAACATCTACGGATACGGCGACAGCCCCGGCGGTGGTTGCGAACTCGGTCCTTCCGCTAAG
GGCGTCTCCTACATCAACACATTCCAGCGCGGCTCGCAGGAGTCCGTCTGGGAGACGGTCACCCAGCCGAC
GTGCGACAACGGCAAGTACGGTGGGGCGCACGGCTACGTCGACCTGTTCATCCAGGGTTCGACGCCGCCGC
AGTGGAAGTACACCGATGCCCCGGACGCCGACGCCCGTGCCGTCCAGGCTGCGTACTGGGCCTACACCTGG
GCATCGGCGCAGGGCAAGGCAAGCGCGATTGCCCCGACGATCGCCAAGGCGAGCCAAACCGGCGACTACC
TGCGGTACTCGCTCTTTGACAAGTACTTCAAGCAGGTCGGCAACTGCTACCCGGCCAGCTCCTGCCCTGGA
GCAACCGGACGCCAGAGCGAGACCTACCTGATCGGCTGGTACTACGCCTGGGGCGGCTCAAGCCAAGGCT
GGGCCTGGCGCATTGGTGACGGCGCCGCGCACTTCGGCTACCAGAATCCGCTTGCCGCGTGGGCGATGTCG
AACGTGACACCGCTCATTCCGCTCTCGCCCACGGCAAAGAGCGACTGGGCGGCGAGCTTGCAGCGCCAGCT
GGAGTTCTACCAGTGGTTGCAATCCGCGGAAGGAGCCATTGCGGGCGGCGCCACCAACAGCTGGAACGGC
AATTACGGGACCCCGCCGGCCGGAGACTCGACCTTCTACGGCATGGCGTACGACTGGGAGCCGGTCTACCA
CGACCCGCCGAGCAACAACTGGTTCGGCTTCCAGGCGTGGTCCATGGAACGGGTTGCCGAGTACTACTACG
TCACCGGCGACCCGAAGGCCAAGGCGCTGCTCGACAAGTGGGTCGCATGGGTGAAGCCGAATGTCACCAC
CGGTGCCTCATGGTCGATTCCGTCGAATTTGTCCTGGAGCGGCCAACCGGATACCTGGAATCCGAGCAACC
CAGGAACGAATGCCAACCTGCACGTGACCATCACGTCGTCCGGGCAGGACGTCGGTGTTGCCGCGGCGCTC
GCGAAGACACTCGAGTACTACGCGGCAAAATCCGGCGATACGGCCTCGCGCGACCTCGCGAAGGGATTGC
TCGACTCCATGTGGAACAACGACCAGGACAGCCTCGGTGTGAGCACACCGGAGACGCGGACCGACTACTCT
CGGTTCACTCAGGTGTACGACCCGACGACTGGTGACGGCCTCTACATCCCGTCGGGTTGGACGGGGACCAT
GCCCAACGGTGACCAAATCAAGCCGGGTGCGACCTTCCTGAGCATCCGGTCCTGGTACACCAAGGATCCGC
AGTGGTCGAAGGTGCAGGCGTACCTCAACGGCGGGCCTGCTCCGACGTTCAACTACCACCGGTTCTGGGCG
GAGTCCGACTTCGCGATGGCGAACGCCGATTTTGGCATGCTCTTCCCATCCGGGTCGCCCAGCCCGACCCC
GAGCCCGACTCCGACGTCGTCCCCGAGCCCGACTCCGAGCAGCTCGCCGACGCCGTCGCCCAGCCCGTCAC
CGACCGGCGACACCACGCCGCCGAGCGTGCCGACGGGTCTTCAGGTCACCGGGACAACGACGTCGTCCGTG
TCGCTCAGCTGGACCGCGTCCACCGACAACGTCGGCGTCGCGCACTACAACGTGTACCGAAACGGCACGCT
GGTGGGTCAGCCGACAGCGACGTCGTTCACGGACACCGGCCTGGCTGCTGGCACGTCGTACACGTACACAG
TGGCGGCCGTTGATGCGGCCGGTAACACGTCGGCGCAGAGCTTCGCCGGTGACAGCGACGACGGCATCGC
CGTCGCGAGCCCGTCGCCGAGCCCGACTCCGACGTCGTCCCCGAGCCCAACGCCGTCGCCGACACCGTCAC
CGACGTCCACCAGCGGCGCATCGTGCACTGCTACCTACGTTGTCAATAGCGACTGGGGTAGCGGCTTCACG
ACAACCGTGACCGTGACGAACACCGGCACCAGGGCCACCAGTGGCTGGACGGTCACGTGGAGCTTTGCCG
GTAATCAGACGGTCACCAACTACTGGAACACCGCGCTGACGCAATCCGGAAAGTCGGTGACCGCAAAGAA
CCTGAGTTACAACAACGTCATCCAACCTGGTCAGTCGACGACCTTTGGATTCAACGGAAGTTACTCAGGAA
CAAACACCGCGCCGACGCTCAGCTGCACGGCAAGCTGA

The Gux1 polynucleotide molecules of the invention are preferably isolated molecules encoding the Gux1 polypetide having an amino acid sequence as shown in Table 1 and SEQ ID NO:1, as well as

derivatives, variants, and useful fragments of the Gux1 polynucleotide. The Gux1 polynucleotide sequence can include deletions, substitutions, or additions to the nucleic acid sequence of Table 2 and SEQ ID NO:2.

The Gux1 polynucleotide molecule of the invention can be cDNA, chemically synthesized DNA, DNA amplified by PCR, RNA, or combinations thereof. Due to the degeneracy of the genetic code, two DNA sequences may differ and yet encode identical amino acid sequences. The present invention thus provides an isolated polynucleotide molecule having a Gux1 nucleic acid sequence encoding Gux1 polypeptide, where the nucleic acid sequenc encodes a polypeptide having the complete amino acid sequences as shown in Table 1 and SEQ ID NO:1 , or variants, derivatives, and fragments thereof.

The Gux1 polynucleotides of the invention have a nucleic acid sequence that is at least about 60% identical to the nucleic acid sequence shown in Table 2 and SEQ ID NO:2, in some embodiments at least about 70% identical to the nucleic acid sequence shown in Table 2 and SEQ ID NO:2, and in other embodiments at least about 90% identical to the nucleic acid sequence shown in Table 2 and SEQ ID NO:2. Nucleic acid sequence identity is determined by known methods, for example by aligning two sequences in a software program such as the BLAST program (Altschul, S.F et al. (1990) J. Mol. Biol. 215:403-410, from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/).

The Gux1 polynucleotide molecules of the invention also include isolated polynucleotide molecules having a nucleic acid sequence that hybridizes under high stringency conditions (as defined above) to a the nucleic acid sequence shown in Table 2 and SEQ ID NO:2. Hybridization of the polynucleotide is to at least about 15 contiguous nucleotides, or at least about 20 contiguous nucleotides, and in other embodiments at least about 30 contiguous nucleotides, and in still other embodiments at least about 100 contiguous nucleotides of the nucleic acid sequence shown in Table 2 and SEQ ID NO:2.

Useful fragments of the Gux1-encoding polynucleotide molecules described herein, include probes and primers. Such probes and primers can be used, for example, in PCR methods to amplify and detect the presence of Gux1 polynucleotides *in vitro*, as well as in Southern and Northern blots for analysis of Gux1. Cells expressing the Gux1 polynucleotide molecules of the invention can also be identified by the use of such probes. Methods for the production and use of such primers and probes are known. For

PCR, 5' and 3' primers corresponding to a region at the termini of the Gux1 polynucleotide molecule can be employed to isolate and amplify the Gux1 polynucleotide using conventional techniques.

Other useful fragments of the Gux1 polynucleotides include antisense or sense oligonucleotides comprising a single-stranded nucleic acid sequence capable of binding to a target Gux1 mRNA (using a sense strand), or DNA (using an antisense strand) sequence.

Vectors and Host Cells:

The present invention also provides vectors containing the polynucleotide molecules of the invention, as well as host cells transformed with such vectors. Any of the polynucleotide molecules of the invention may be contained in a vector, which generally includes a selectable marker and an origin of replication, for propagation in a host. The vectors further include suitable transcriptional or translational regulatory sequences, such as those derived from a mammalian, microbial, viral, or insect genes, operably linked to the Gux1 polynucleotide molecule. Examples of such regulatory sequences include transcriptional promoters, operators, or enhancers, mRNA ribosomal binding sites, and appropriate sequences which control transcription and translation. Nucleotide sequences are operably linked when the regulatory sequence functionally relates to the DNA encoding the target protein. Thus, a promoter nucleotide sequence is operably linked to a Gux1 DNA sequence if the promoter nucleotide sequence directs the transcription of the Gux1 sequence.

Selection of suitable vectors for the cloning of Gux1 polynucleotide molecules encoding the target Gux1 polypeptides of this invention will depend upon the host cell in which the vector will be transformed, and, where applicable, the host cell from which the target polypeptide is to be expressed. Suitable host cells for expression of Gux1 polypeptides include prokaryotes, yeast, and higher eukaryotic cells, each of which is discussed below.

The Gux1 polypeptides to be expressed in such host cells may also be fusion proteins that include regions from heterologous proteins. As discussed above, such regions may be included to allow, for example, secretion, improved stability, or facilitated purification of the Gux1 polypeptide. For example, a nucleic acid sequence encoding an appropriate signal peptide can be incorporated into an expression vector. A nucleic acid sequence encoding a signal peptide (secretory leader) may be fused in-frame to the Gux1 sequence so that Gux1 is translated as a fusion protein comprising the signal

peptide. A signal peptide that is functional in the intended host cell promotes extracellular secretion of the Gux1 polypeptide. Preferably, the signal sequence will be cleaved from the Gux1 polypeptide upon secretion of Gux1 from the cell. Non-limiting examples of signal sequences that can be used in practicing the invention include the yeast I-factor and the honeybee melatin leader in Sf9 insect cells.

Suitable host cells for expression of target polypeptides of the invention include prokaryotes, yeast, and higher eukaryotic cells. Suitable prokaryotic hosts to be used for the expression of these polypeptides include bacteria of the genera *Escherichia, Bacillus, and Salmonella,* as well as members of the genera *Pseudomonas, Streptomyces,* and *Staphylococcus.* For expression in prokaryotic cells, for example, in *E. coli*, the polynucleotide molecule encoding Gux1 polypeptide preferably includes an N-terminal methionine residue to facilitate expression of the recombinant polypeptide. The N-terminal Met may optionally be cleaved from the expressed polypeptide.

Expression vectors for use in prokaryotic hosts generally comprise one or more phenotypic selectable marker genes. Such genes encode, for example, a protein that confers antibiotic resistance or that supplies an auxotrophic requirement. A wide variety of such vectors are readily available from commercial sources. Examples include pSPORT vectors, pGEM vectors (Promega, Madison, WI), pPROEX vectors (LTI, Bethesda, MD), Bluescript vectors (Stratagene), and pQE vectors (Qiagen).

Gux1 can also be expressed in yeast host cells from genera including *Saccharomyces, Pichia,* and *Kluveromyces.* Preferred yeast hosts are *S. cerevisiae* and *P. pastoris.* Yeast vectors will often contain an origin of replication sequence from a 2T yeast plasmid, an autonomously replicating sequence (ARS), a promoter region, sequences for polyadenylation, sequences for transcription termination, and a selectable marker gene. Vectors replicable in both yeast and *E. coli* (termed shuttle vectors) may also be used. In addition to the above-mentioned features of yeast vectors, a shuttle vector will also include sequences for replication and selection in *E. coli.* Direct secretion of the target polypeptides expressed in yeast hosts may be accomplished by the inclusion of nucleotide sequence encoding the yeast I-factor leader sequence at the 5' end of the Gux1-encoding nucleotide sequence.

Insect host cell culture systems can also be used for the expression of Gux1 polypeptides. The target polypeptides of the invention are preferably expressed using a baculovirus expression system, as described, for example, in the review by Luckow and Summers, 1988 *Bio/Technology* 6:47.

The choice of a suitable expression vector for expression of Gux1 polypeptides of the invention will depend upon the host cell to be used. Examples of suitable expression vectors for *E. coli* include pET, pUC, and similar vectors as is known in the art. Preferred vectors for expression of the Gux1 polypeptides include the shuttle plasmid pIJ702 for *Streptomyces lividans*, pGAPZalpha-A, B, C and pPICZalpha-A, B, C (Invitrogen) for *Pichia pastoris*, and pFE-1 and pFE-2 for filamentous fungi and similar vectors as is known in the art.

Modification of a Gux1 polynucleotide molecule to facilitate insertion into a particular vector (for example, by modifiying restriction sites), ease of use in a particular expression system or host (for example, using preferred host codons), and the like, are known and are contemplated for use in the invention. Genetic engineering methods for the production of Gux1 polypeptides include the expression of the polynucleotide molecules in cell free expression systems, in cellular hosts, in tissues, and in animal models, according to known methods.

Compositions

The invention provides compositions containing a substantially purified Gux1 polypeptide of the invention and an acceptable carrier. Such compositions are administered to biomass, for example, to degrade the cellulose in the biomass into simpler carbohydrate units and ultimately, to sugars. These released sugars from the cellulose are converted into ethanol by any number of different catalysts. Such compositions may also be included in detergents for removal, for example, of cellulose containing stains within fabrics, or compositions used in the pulp and paper industry, to address conditions associated with cellulose content. Compositions of the present invention can be used in stonewashing jeans such as is well known in the art. Compositions can be used in the biopolishing of cellulosic fabrics, such as cotton, linen, rayon and Lyocell.

The invention provides pharmaceutical compositions containing a substantially purified Gux1 polypeptide of the invention and if necessary a pharmaceutically acceptable carrier. Such pharmaceutical compositions are administered to cells, tissues, or patients, for example, to aid in delivery or targeting of other pharmaceutical compositions. For example, Gux1 polypeptides may be used where carbohydrate-mediated liposomal interactions are involved with target cells. Vyas SP et al. (2001), J. Pharmacy & Pharmaceutical Sciences May-Aug 4(2): 138-58.

The invention also provides reagents, compositions, and methods that are useful for analysis of Gux1 activity and for the analysis of cellulose breakdown.

Compositions of the present invention may also include other known cellulases, and preferably, other known thermal tolerant cellulases for enhanced treatment of cellulose.

Antibodies

The polypeptides of the present invention, in whole or in part, may be used to raise polyclonal and monoclonal antibodies that are useful in purifying Gux1, or detecting Gux1 polypeptide expression, as well as a reagent tool for characterizing the molecular actions of the Gux1 polypeptide. Preferably, a peptide containing a unique epitope of the Gux1 polypeptide is used in preparation of antibodies, using conventional techniques. Methods for the selection of peptide epitopes and production of antibodies are known. See, for example, *Antibodies: A Laboratory Manual*, Harlow and Land (eds.), 1988 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.; *Monoclonal Antibodies, Hybridomas: A New Dimension in Biological Analyses*, Kennet *et al.* (eds.), 1980 Plenum Press, New York.

Assays

Agents that modify, for example, increase or decrease, Gux1 hydrolysis or degradation of cellulose can be identified, for example, by assay of Gux1 cellulase activity and/or analysis of Gux1 binding to a cellulose substrate. Incubation of cellulose in the presence of Gux1 and in the presence or absence of a test agent and correlation of cellulase activity or carbohydrate binding permits screening of such agents. For example, cellulase activity and binding assays may be performed in a manner similar to those described in Irwin et al., J. Bacteriology 180(7): 1709-1714 (April 1998).

The Gux1 stimulated activity is determined in the presence and absence of a test agent and then compared. A lower Gux1 activated test activity in the presence of the test agent, than in the absence of the test agent, indicates that the test agent has decreased the activity of the Gux1. A higher Gux1 activated test activity in the presence of the test agent than in the absence of the test agent indicates that the test agent has increased the activity of the Gux1. Stimulators and inhibitors of Gux1 may be used to augment, inhibit, or modify Gux1 mediated activity, and therefore may have potential industrial uses as well as potential use in the further elucidation of Gux1's molecular actions.

Therapeutic Applications

The Gux1 polypeptides of the invention are effective in adding in delivery or targeting of other pharmaceutical compositions within a host. For example, Gux1 polypeptides may be used where carbohydrate-mediated liposomal interactions are involved with target cells. Vyas SP et al. (2001), *J. Pharm Pharm Sci* May-Aug 4(2): 138-58.

Gux1 polynucleotides and polypeptides, including vectors expressing Gux1, of the invention can be formulated as pharmaceutical compositions and administered to a host, preferably mammalian host, including a human patient, in a variety of forms adapted to the chosen route of administration. The compounds are preferably administered in combination with a pharmaceutically acceptable carrier, and may be combined with or conjugated to specific delivery agents, including targeting antibodies and/or cytokines.

Gux1 can be administered by known techniques, such as orally, parentally (including subcutaneous injection, intravenous, intramuscular, intrasternal or infusion techniques), by inhalation spray, topically, by absorption through a mucous membrane, or rectally, in dosage unit formulations containing conventional non-toxic pharmaceutically acceptable carriers, adjuvants or vehicles. Pharmaceutical compositions of the invention can be in the form of suspensions or tablets suitable for oral administration, nasal sprays, creams, sterile injectable preparations, such as sterile injectable aqueous or oleagenous suspensions or suppositories.

For oral administration as a suspension, the compositions can be prepared according to techniques well-known in the art of pharmaceutical formulation. The compositions can contain microcrystalline cellulose for imparting bulk, alginic acid or sodium alginate as a suspending agent, methylcellulose as a viscosity enhancer, and sweeteners or flavoring agents. As immediate release tablets, the compositions can contain microcrystalline cellulose, starch, magnesium stearate and lactose or other excipients, binders, extenders, disintegrants, diluents and lubricants known in the art.

For administration by inhalation or aerosol, the compositions can be prepared according to techniques well-known in the art of pharmaceutical formulation. The compositions can be prepared

as solutions in saline, using benzyl alcohol or other suitable preservatives, absorption promoters to enhance bioavailability, fluorocarbons or other solubilizing or dispersing agents known in the art.

For administration as injectable solutions or suspensions, the compositions can be formulated according to techniques well-known in the art, using suitable dispersing or wetting and suspending agents, such as sterile oils, including synthetic mono- or diglycerides, and fatty acids, including oleic acid.

For rectal administration as suppositories, the compositions can be prepared by mixing with a suitable non-irritating excipient, such as cocoa butter, synthetic glyceride esters or polyethylene glycols, which are solid at ambient temperatures, but liquefy or dissolve in the rectal cavity to release the drug.

Preferred administration routes include orally, parenterally, as well as intravenous, intramuscular or subcutaneous routes. More preferably, the compounds of the present invention are administered parenterally, i.e., intravenously or intraperitoneally, by infusion or injection.

Solutions or suspensions of the compounds can be prepared in water, isotonic saline (PBS) and optionally mixed with a nontoxic surfactant. Dispersions may also be prepared in glycerol, liquid polyethylene, glycols, DNA, vegetable oils, triacetin and mixtures thereof. Under ordinary conditions of storage and use, these preparations may contain a preservative to prevent the growth of microorganisms.

The pharmaceutical dosage form suitable for injection or infusion use can include sterile, aqueous solutions or dispersions or sterile powders comprising an active ingredient which are adapted for the extemporaneous preparation of sterile injectable or infusible solutions or dispersions. In all cases, the ultimate dosage form should be sterile, fluid and stable under the conditions of manufacture and storage. The liquid carrier or vehicle can be a solvent or liquid dispersion medium comprising, for example, water, ethanol, a polyol such as glycerol, propylene glycol, or liquid polyethylene glycols and the like, vegetable oils, nontoxic glyceryl esters, and suitable mixtures thereof. The proper fluidity can be maintained, for example, by the formation of liposomes, by the maintenance of the required particle size, in the case of dispersion, or by the use of nontoxic surfactants. The prevention

of the action of microorganisms can be accomplished by various antibacterial and antifungal agents, for example, parabens, chlorobutanol, phenol, sorbic acid, thimerosal, and the like. In many cases, it will be desirable to include isotonic agents, for example, sugars, buffers, or sodium chloride. Prolonged absorption of the injectable compositions can be brought about by the inclusion in the composition of agents delaying absorption--for example, aluminum monosterate hydrogels and gelatin.

Sterile injectable solutions are prepared by incorporating the compounds in the required amount in the appropriate solvent with various other ingredients as enumerated above and, as required, followed by filter sterilization. In the case of sterile powders for the preparation of sterile injectable solutions, the preferred methods of preparation are vacuum drying and freeze-drying techniques, which yield a powder of the active ingredient plus any additional desired ingredient present in the previously sterile-filtered solutions.

Industrial Applications

The Gux1 polypeptides of the invention are effective cellulases. In the methods of the invention, the cellulose degrading effects of Gux1 are achieved by treating biomass at a ratio of about 1 to about 50 of Gux1:biomass. Gux1 may be used under extreme conditions, for example, elevated temperatures and acidic pH. Treated biomass is degraded into simpler forms of carbohydrates, and in some cases glucose, which is then used in the formation of ethanol or other industrial chemicals, as is known in the art. Other methods are envisioned to be within the scope of the present invention, including methods for treating fabrics to remove cellulose-containing stains and other methods already discussed. Gux1 polypeptides can be used in any known application currently utilizing a cellulase, all of which are within the scope of the present invention.

Having generally described the invention, the same will be more readily understood by reference to the following examples, which are provided by way of illustration and are not intended as limiting.

EXAMPLES

Example 1: Molecular Cloning of Gux1

Genomic DNA was isolated from *Acidothermus cellulolyticus* and purified by banding on cesium chloride gradients. Genomic DNA was partially digested with Sau 3A and separated on agarose gels. DNA fragments in the range of 9-20 kilobase pairs were isolated from the gels. This purified Sau 3A digested genomic DNA was ligated into the Bam H1 acceptor site of purified EMBL3 lambda phage arms (Clontech, San Diego, Calif.). Phage DNA was packaged according to the manufacturer's specification and plated with *E. Coli* LE392 in top agar which contained the soluble cellulose analog, carboxymethylcellulose (CMC). The plates were incubated overnight (12-24 hours) to allow transfection, bacterial growth, and plaque formation. Plates were stained with Congo Red followed by destaining with 1 M NaCl. Lambda plaques harboring endoglucanase clones showed up as unstained plaques on a red background.

Lambda clones which screened positive on CMC-Congo Red plates were purified by successive rounds of picking, plating and screening. Individual phage isolates were named SL-1, SL-2, SL-3, and SL-4. Subsequent subcloning efforts employed the SL-3 clone which contained an approximately 14.2 kilobase fragment of *Acidothermus cellulolyticus* genomic DNA.

Template DNA was constructed using a 9 kilobase Bam H1 fragment obtained from the 14.2 kilobase lambda clone SL-3 prepared from *Acidothermus cellulolyticus* genomic DNA. The 9 kilobase Bam H1 fragment from SL-3 was subcloned into pDR540 to generate a plasmid NREL501. NREL501 was sequenced by the primer walking method as is known in the art. NREL501 was then subcloned into pUC19 using restriction enzymes Pst I and Eco RI and transformed into *E. coli* XL1-blue (Stratagene) for the production of template DNA for sequencing. Each subclone was sequenced from both the forward and reverse directions. DNA for sequencing was prepared from an overnight growth in 500 mL LB broth using a megaprep DNA purification kit from Promega. The templated DNA was PEG precipitated and suspended in de-ionized water and adjusted to a final concentration of 0.25 milligrams/mL.

Custom primers were designed by reading upstream known sequence and selecting segments of an appropriate length to function, as is well known in the art. Primers for cycle sequencing were

synthesized at the Macromolecular Resources Facility located at Colorado State University in Fort Collins , Colorado. Typically the sequencing primers were 26 to 30 nucleotides in length, but were sometimes longer or shorter to accommodate a melting temperature appropriate for cycle sequencing. The sequencing primers were diluted in de-ionized water, the concentration measured using UV absorbance at 260 nm, and then adjusted to a final concentration of 5 pmol/microL.

Templates and sequencing primers were shipped to the Iowa State University DNA Sequencing Facility at Ames, Iowa for sequencing using standard chemistries for cycle sequencing. In some cases, regions of the template that sequenced poorly using the standard protocols and dye terminators were repeated with the addition of 2 microL DMSO and by using nucleotides optimized for the sequencing of high GC content DNA.

Sequencing data from primer walking and subclones were assembled together to verify that all SL-3 regions had been sequenced from both strands. An open reading frame (ORF) was found in the 9 kilobase Bam H1 fragment, C-terminal of E1 (U.S. Patent 5,536,655), termed Gux1. An ORF of 3366 bp [SEQ ID NO:2] and deduced amino acid sequence [SEQ ID NO:1] are shown in Tables 1 and 2. The amino acid sequence predicted by SEQ ID NO:1 was determined to have significant homology to known cellulases, as is shown below in Example 2 and Table 3.

The amino acid sequence represents a novel member of the family of proteins with cellulase activity. Due to the source of isolation, from the thermophilic *Acidothermus cellulolyticus*, Gux1 is a novel member of cellulases with properties including thermal tolerance. It is also known that thermal tolerant enzymes may have other properties (see definition above).

Example 2: Gux1 includes a GH48 catalytic domain

Sequence alignments and comparisons of the amino acid sequences of the *Acidothermus cellulolyticus* Gux1 catalytic domain (approximately amino acids 231 to 870), *Cellulomonas fimi* (cellobiohydrolase B) and *Thermobifida fusca* (exocellulase E6) polypeptides were prepared, using the ClustalW program (Thompson J.D et al. (1994), Nucleic Acids Res. 22:4673-4680 from EMBL European Bioinformatics Institute website (http://www.ebi.ac.uk/)). An examination of the amino acid sequence alignment of the GH48 domains indicates that the amino acid sequence of Gux1 catalytic domain is homologous to the amino acid sequences of known GH48 family catalytic

domains for *C. fimi* cellobiohydrolase B and *T. fusca* exocellulase E6 (see Table 3). In Table 3, the notations are as follows: an asterisk "*" indicates identical or conserved residues in all sequences in the alignment; a colon ":" indicates conserved substitutions; a period "." indicates semi-conserved substitutions; and a hyphen "-" indicates a gap in the sequence. The amino acid sequence predicted for the Gux1 GH48 domain is approximately 64 % identical to the *C. fimi* cellobiohydrolase B GH48 domain and approximately 57 % identical to the *T. fusca* exocellulase E6 GH48 domain, indicating that the Gux1 catalytic domain is a member of the GH48 family (Henrissat et al., (1991) *supra*).

Table 3. Multiple amino acid sequence alignment of a Gux1 first catalytic domain and polypeptides with Glycoside Hydrolase Family 48 catalytic domains.

Multialignment of related Glycoside Hydrolase Family 48 catalytic domain
GH48_Ace: *Acidothermus cellulolyticus* Gux1 catalytic domain GH48
GuxB_Cfi: Cellulomonas fimi CBHB(beta-1,4-exocellobiohydrolase). GeneBank Acc. # AAB00822
E6_Tfu: Thermobifida fusca E6 (beta-1,4-exocellulase). GeneBank Acc. # AF144563

```
GH48_Ace        PYIQRFLTMYNKIHDPANGYFSPQG---IPYHSVETLIVEAPDYGHETTSEAYSFWLWLEATY
GuxB_Cfi        EYAQRFLAQYDKIKDPANGYFSAQG---IPYHAVETLMVEAPDYGHETTSEAYSYWLWLEALY
E6_Tfu          SYDQAFLEQYEKIKDPASGYFREFNGLLVPYHSVETMIVEAPDHGHQTTSEAFSYYLWLEAYY
                 * * **  *:**:***.***    .   :***:***::*****:**:*****:*::***** *

GH48_Ace        GAVTGNWTPFNNAWTTMETYMIPQHADQPNNASYNPNSPASYAPEEPLPSMYPVAIDSSV
GuxB_Cfi        GQVTQDWAPLNHAWDTMEKYMIPQSVDQPTNSFYNPNSPATYAPEFNHPSSYPSQLNSGI
E6_Tfu          GRVTGDWKPLHDAWESMETFIIPGTKDQPTNSAYNPNSPATYIPEQPNADGYPSPLMNNV
                * ** :* *::.** :**.::**   ***.*: *******:* **   .. **  : ..:

GH48_Ace        PVGHDPLAAELQSTYGTPDIYGMHWLADVDNIYGYGDSPGGGCELGPSAKGVSYINTFQR
GuxB_Cfi        SGGTDPIGAELKATYGNADVYQMHWLADVDNIYGFGATPGAGCTLGPTATGTSFINTFQR
E6_Tfu          PVGQDPLAQELSSTYGTNEIYGMHWLLDVDNVYGFGFCGDG------TDDAPAYINTYQR
                . * **:. **.:***. ::* **** ****:**:*   ..      :  . ::***:**

GH48_Ace        GSQESVWETVTQPTCDNGKYGGAHGYVDLFIQG-STPPQWKYTDAPDADARAVQAAYWAY
GuxB_Cfi        GPQESVWETVPQPSCEEFKYGGKNGYLDLFTKDASYAKQWKYTSASDADARAVEAVYWAN
E6_Tfu          GARESVWETIPHPSCDDFTHGGPNGYLDLFTDDQNYAKQWRYTNAPDADARAVQVMFWAH
                *.:******:.:*:*:: .:** :**:*** .. . . **:**.*.*******:. :**

GH48_Ace        TWASAQGKASAIAPTIAKASQTGDYLRYSLFDKYFKQVGNCYPASSCPGATGRQSETYLI
GuxB_Cfi        QWATEQGKAADVAATVAKAAKMGDYLRYTLFDKYFKKIG--CTSPTCAAGQGREAAHYLL
E6_Tfu          EWAKEQGKENEIAGLMDKASKMGDYLRYAMFDKYFKKIGNCVGATSCPGGQGKDSAHYLL
                 **. ***   :*  : **:: ******::******::*    :.:*... *:::  **:

GH48_Ace        GWYYAWGGS---SQGWAWRIGDGAAHFGYQNPLAAWAMSNVTPLIPLSPTAKSDWAASLQ
GuxB_Cfi        SWYMAWGGATDTSSGWAWRIGSSHAHFGYQNPLAAWALSTDPKLTPKSPTAKADWAASMQ
E6_Tfu          SWYYSWGGSLDTSSAWAWRIGSSSSHQGYQNVLAAYALSQVPELQPDSPTGVQDWATSFD
                .** :***:   *..******.. :* **** ***:*:*  . * * ***.  ***:*::

GH48_Ace        RQLEFYQWLQSAEGAIAGGATNSWNGNYGTPPAGDSTFYGMAYDWEPVYHDPPSNNWFGF
GuxB_Cfi        RQLEFYTWLQASNGGIAGGATNSWDGAYAQPPAGTPTFYGMGYTEAPVYVDPPSNRWFGM
E6_Tfu          RQLEFLQWLQSAEGGIAGGATNSWKGSYDTPPTGLSQFYGMYYDWQPVWNDPPSNNWFGF
                ***** ***:::*.*********.* *  **:* . **** *   **: *****.***:

GH48_Ace        QAWSMERVAEYYYVTGDPKAKALLDKWVAWVKPNVTTG-----ASWSIPSNLSWSGQPDT
GuxB_Cfi        QAWGVQRVAELYYASGNAQAKKILDKWVPWVVANISTDG----ASWKVPSELKWTGKPDT
E6_Tfu          QVWNMERVAQLYYVTGDARAEAILDKWVPWAIQHTDVDADNGGQNFQVPSDLEWSGQPDT
                *.*.::***: **.:*:.:*: :*****.*.  :  ..       .:.:**:*.*:*:***

GH48_Ace        WNPSNPGTNANLHVTITSSGQDVGVAAALAKTLEYYAAKSGDTASRDLAKGLLDSMWNND
GuxB_Cfi        WNAAAPTGNPGLTVEVTSYGQDVGVAADTARALLFYAAKSGDTASRDKAKALLDAIWANN
E6_Tfu          WTG-TYTGNPNLHVQVVSYSQDVGVTAALAKTLMYYAKRSGDTTALATAEGLLDALLAHR
                *.      *..* * :.* .*****:*  *::* :** :****::   *:.***::  :

GH48_Ace        QDSLGVSTPETRTDYSRFTQVYDPTTGDGLYIPSGWTGTMPNGDQIKPGATFLSIRSWYT
GuxB_Cfi        QDPLGVSAVETRGDYKRFDDTYVAN-GDGIYIPSGWTGTMPNGDVIKPGVSFLDIRSFYK
E6_Tfu          -DSIGIATPEQ-PSWDRLDDPWDGS--EGLYVPPGWSGTMPNGDRIEPGATFLSIRSFYK
                 *.:*::: *   .:.*: : :  .  :*:*:*.**:******* *:**.:**.***:*.

GH48_Ace        KDPQWSKVQAYLNGG---PAPTFNYHRFWAESDFAMANADFGMLFPSGSP
GuxB_Cfi        KDPNWSKVQTFLDGG---AEPQFRYHRFWAQTAVAGALADYARLFDDGTT
E6_Tfu          NDPLWPQVEAHLNDPQNVPAPIVERHRFWAQVEIATAFAAHDELFGAGAP
                :** *.:*::.*:.    . * .. *****:  .* * * .  **  *:.
```

Example 3: Mixed Domain GH48, CBD II, CBD III Genes and Hybrid Polypeptides

From the putative locations of the domains in the Gux1 cellulase sequence given above and in comparable cloned cellulase sequences from other species, one can separate individual domains and combine them with one or more domains from different sequences. The significant similarity

between cellulase genes permit one by recombinant techniques to arrange one or more domains from the *Acidothermus cellulolyticus* Gux1 cellulase gene with one or more domains from a cellulase gene from one or more other microorganisms. Other representative endoglucanase genes include *Bacillus polymyxa* beta-(1,4) endoglucanase (Baird et al, Journal of Bacteriology, 172: 1576-86 (1992)) and *Xanthomonas campestris* beta-(1,4)-endoglucanase A (Gough et al, Gene 89:53-59 (1990)). The result of the fusion of any two or more domains will, upon expression, be a hybrid polypeptide. Such hybrid polypeptides can have one or more catalytic or binding domains. For ease of manipulation, recombinant techniques may be employed such as the addition of restriction enzyme sites by site-specific mutagenesis. If one is not using one domain of a particular gene, any number of any type of change including complete deletion may be made in the unused domain for convenience of manipulation.

It is understood for purposes of this disclosure, that various changes and modifications may be made to the invention that are well within the scope of the invention. Numerous other changes may be made which will readily suggest themselves to those skilled in the art and which are encompassed in the spirit of the invention disclosed herein and as defined in the appended claims.

This specification contains numerous citations to references such as patents, patent applications, and publications. Each is hereby incorporated by reference for all purposes.